FAST-HASOC 2025: Multimodal and Multilingual Approaches for Hate Speech and Offensive Content Detection in Hindi Memes

Muhammad Rafi[1,*], Saif Ur Rehman Awan[2,†], Ramsha Jat[3,†], Aiman Falak[4,†], Fatimah Ansari[5,†], Ahmed Raza[6,†] and Sagar Chabbriya[7,†]

[1]National University of Computer and Emerging Sciences, Islamabad, Pakistan

Abstract

This paper describes our **Team FAST's** participation in the HASOC 2025 shared task on offensive content detection in Hindi memes. The dataset consists of multimodal samples with OCR-extracted text and raw images, annotated across four subtasks: sentiment, sarcasm, vulgarity, and abuse detection. We propose a multimodal framework that integrates classical machine learning and deep learning models. Our contributions are as follows: (i) a tailored preprocessing pipeline for noisy OCR and Hindi, English code-mixing using curated stopword and vulgar dictionaries, (ii) a combination of lightweight classical models (TF–IDF + Random Forest) with neural approaches (CNN, BiLSTM, ResNet50). Our system achieved its best performance in **Vulgarity detection with a Macro-F1 score of 0.75**. Code, data splits, and preprocessing resources are available at: https://github.com/fatimahansari/hindi-HASOC-2025.

Keywords

Hate Speech Detection, Multimodal NLP, Hindi Memes, TF–IDF, Random Forest, CNN, BiLSTM, ResNet50, HASOC 2025

1. Introduction

Offensive content on social media spans textual and visual modalities, often with subtle cues, code-mixing, and transliteration. The HASOC 2025 shared task [1] focuses on Hindi memes, combining OCR-extracted text and meme images, with subtasks in sentiment, sarcasm, vulgarity, and abuse detection. These challenges are amplified by noisy OCR, informal Hinglish, and implicit insults. We present our approach, which combines both classical and neural models under a multimodal pipeline. Unlike prior editions focusing on monomodal text [2], our system explicitly fuses image and text features for vulgarity and incorporates Hindi-specific lexicons for preprocessing. Our contributions are:

- A preprocessing framework for noisy OCR and Hindi–English code-mixing.
- A multimodal architecture combining TF–IDF + Random Forest with ResNet50.

2. Dataset and Resources

We used the HASOC 2025 Hindi meme dataset:

- **Train:** 1133 samples with labels.
- **Test:** 767 samples.

Each sample includes an image and OCR text annotated for four subtasks. Additional resources:

- `hindi_stopwords.json` — curated Hindi/Hinglish stopword list.

Forum for Information Retrieval Evaluation, December 17-20, 2025, Varanasi, India

*Corresponding author.

†These authors contributed equally.

muhammad.rafi@nu.edu.pk (M. Rafi); saifurrehman@nu.edu.pk (S. U. R. Awan); ramsha.jat@nu.edu.pk (R. Jat); aiman.falak@nu.edu.pk (A. Falak); fatimahansari614@gmail.com (F. Ansari); ahmedraza9332@gmail.com (A. Raza); sagarchhabriya34@gmail.com (S. Chabbriya)

- `hindi-offensive-words-original.json` — offensive lexicon mapped to neutral terms.

These resources are publicly released with our code repository.

3. System Architecture

Figure 1 shows the system pipeline: preprocessed text is passed through task-specific models; images are processed for vulgarity detection and fused with text predictions.

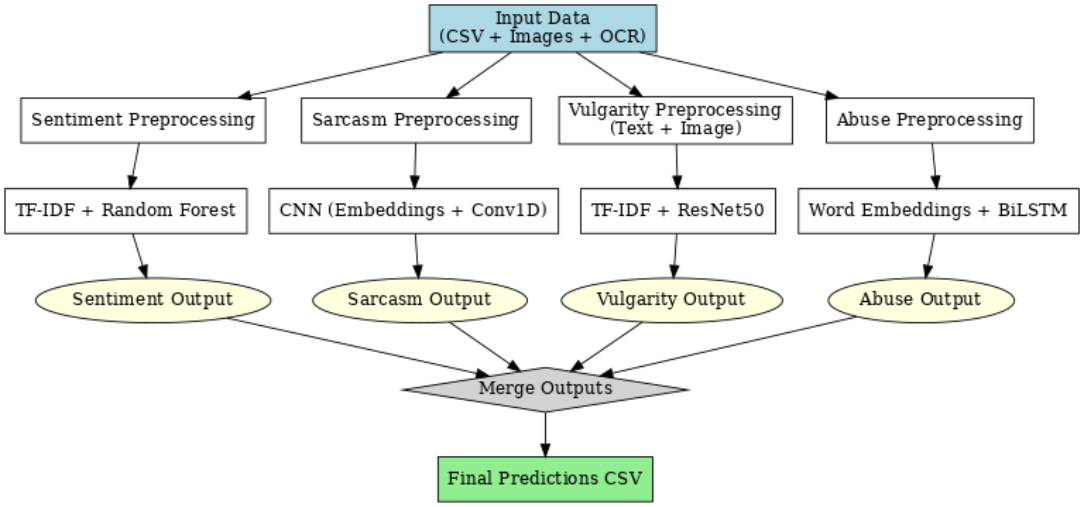

Figure 1: Overall multimodal system pipeline for HASOC 2025 subtasks.

3.1. Preprocessing

Our text preprocessing includes:

1. Cleaning URLs, emails, non-Devanagari symbols.
2. Stopword removal (Hindi/Hinglish).
3. Offensive word replacement using the vulgar dictionary.
4. Tokenization and language-aware normalization.

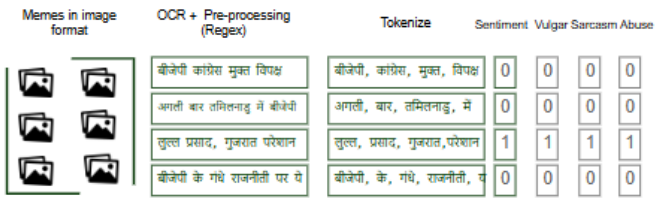

Figure 2: Pre-processing the OCR text

3.2. Models

Sentiment: For sentiment detection, we chose a combination of TF–IDF features and a Random Forest classifier. TF–IDF is effective in representing textual data from short social media posts and OCR

text because it captures the importance of words and bigrams while ignoring overly frequent stopwords. Random Forest, being an ensemble of decision trees, provides robustness against noisy data and works well with sparse, high-dimensional inputs. This model was selected because sentiment cues in Hindi memes are often expressed through explicit keywords or short phrases, making classical feature-based approaches suitable. Additionally, Random Forests handle class imbalance relatively well and offer interpretability compared to deep models.

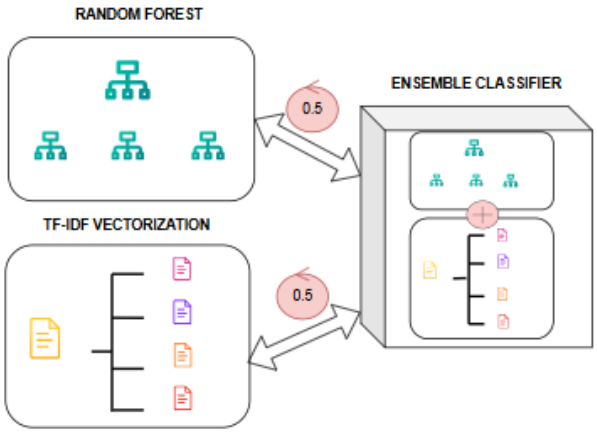

Figure 3: Sentiment detection model using TF–IDF vectorization and Random Forest.

Sarcasm: Sarcasm is typically expressed through subtle lexical patterns, wordplay, and local context within a short sequence of text. To model this, we employed a Convolutional Neural Network (CNN) with an embedding layer, 1D convolutional filters, and global max pooling. The CNN captures n-gram level features by sliding filters over word embeddings, allowing it to learn important combinations of words that signal sarcasm. This architecture is lightweight compared to transformers but effective for short-text sarcasm detection, which often relies on key sarcastic cues rather than long-range dependencies. We chose CNNs because they generalize well on small datasets, train faster, and are less prone to overfitting than more complex models.

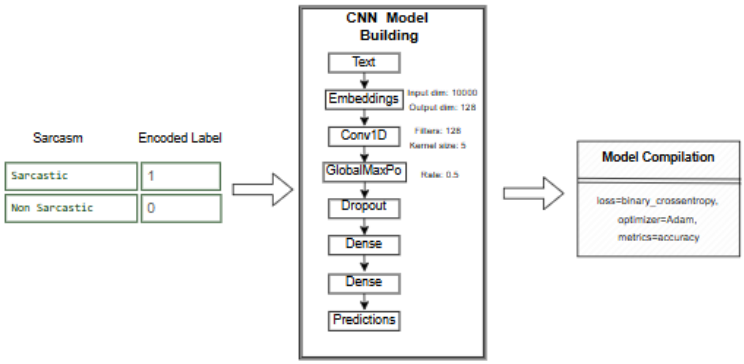

Figure 4: Sarcasm detection model using CNNs.

Vulgarity: This multimodal setup allows the system to leverage complementary strengths, while textual features capture linguistic vulgarity, image embeddings contribute crucial context when offensiveness is implied visually rather than verbally. The decision-level fusion also provides robustness, as errors in one modality can be compensated by the other, leading to more stable and consistent predictions across a wide variety of meme formats. Furthermore, the use of a relatively lightweight architecture like ResNet50 ensures faster inference and reduced computational overhead, making the approach more practical for real world deployment where large volumes of memes need to be processed efficiently. This balance between accuracy, efficiency, and adaptability is particularly important for social media platforms, where offensive content spreads rapidly, and automated systems must detect problematic material in near real-time while maintaining scalability across millions of daily uploads. In addition, this framework supports better generalization across unseen content, as both modalities provide complementary cues that reduce reliance on any single visual or linguistic pattern. The fusion mechanism also improves resilience against adversarial modifications, such as text masking or subtle image manipulation, which are commonly used to dodge detection. By integrating context from both channels, the model can more accurately differentiate between humor and truly vulgar intent, reducing false positives that undermine user trust. Overall, the multimodal design strengthens the system's ability to adapt to evolving forms of online vulgarity, ensuring consistent performance as meme culture and offensive patterns continue to shift over time.

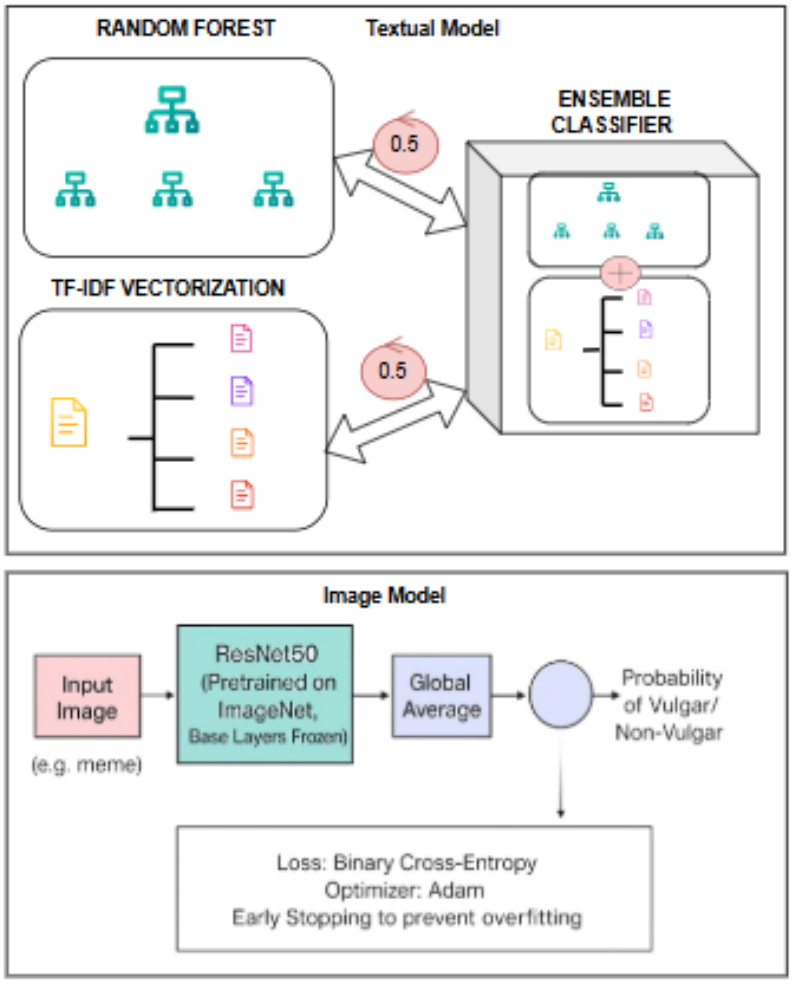

Figure 5: Multimodal vulgarity detection model combining text (TF–IDF + RF) and image (ResNet50) features.

Abuse: Abuse detection is more complex than sentiment or vulgarity because abusive language is often indirect, context-dependent, and highly code-mixed, especially in online spaces where users frequently switch between Hindi and English. To capture sequential dependencies, we employed a Bidirectional LSTM, which processes text in both forward and backward directions, thereby modeling long range dependencies and subtle cues that a unidirectional model might overlook. This helps the model understand relationships between abusive terms and their surrounding context, such as sarcasm or implicit threats. We initialized the model with FastText embeddings trained on Hindi, which provide rich semantic representations even for rare and morphologically complex words, and also adapt well to spelling variations and colloquial usage. Additionally, we incorporated a binary lexicon feature indicating the presence of offensive terms, ensuring that explicit abuse was directly flagged rather than relying solely on contextual embeddings. This hybrid approach was chosen because BiLSTMs effectively capture sequential patterns in short, noisy social media text, while lexicon features act as a safety net for explicit slurs that embeddings might underrepresent. Together, these components create a more comprehensive detection system capable of handling both overt insults and subtle, context-driven abuse.

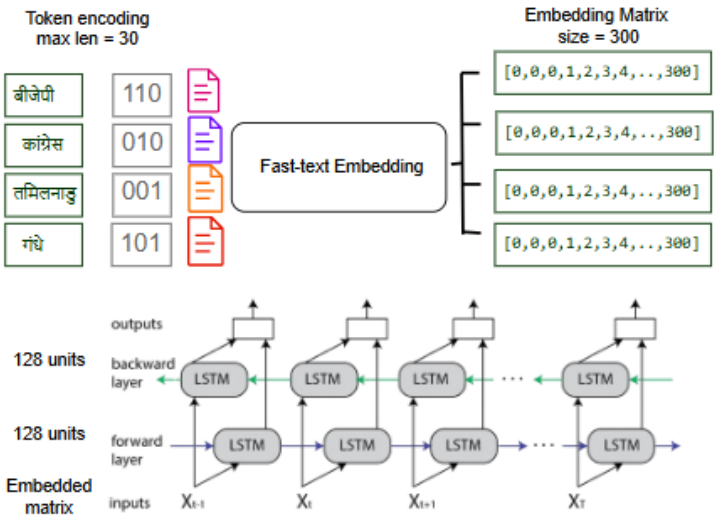

Figure 6: BiLSTM model for abuse detection with FastText embeddings and lexicon features.

Why not Transformers? Although transformer-based models such as BERT and mBERT have shown strong results in offensive language detection, we deliberately chose not to rely on them as our primary models in this work. There are three key reasons. First, the HASOC 2025 dataset for Hindi memes is relatively small (just over 1100 training examples), which makes fine-tuning large transformer models prone to overfitting [3]. In contrast, lighter models such as TF–IDF + Random Forest or CNNs are more data-efficient and generalize better in low-resource settings. Second, OCR-extracted Hindi text is noisy and often code-mixed, with Romanized tokens that pretrained multilingual transformers do not handle well without extensive normalization. Classical models and BiLSTMs with FastText embeddings proved more robust under these conditions, especially when augmented with curated lexicons [4]. Finally, computational efficiency was an important consideration: Random Forests, CNNs, and BiLSTMs are significantly faster to train and deploy, making them practical for iterative experimentation and real-world applications where resources are limited. While transformers remain a promising direction, in this task we prioritized interpretability, efficiency, and robustness in noisy, under-resourced data conditions.

4. Experiments

4.1. Setup

Implemented in Python (scikit-learn, TensorFlow/Keras). Training used 5-fold stratified CV. Hyperparameters were tuned empirically. Experiments were run on a single TPU.

4.2. Baselines

- Majority class prediction.
- Logistic Regression + TF–IDF (text-only).

5. Results

Our models outperform baselines across all subtasks (Table 1). Vulgarity detection particularly benefited from multimodal fusion.

Table 1
HASOC 2025 Hindi memes: test set performance (macro-averaged).

Subtask	Precision	Recall	F1
Sentiment	0.71	0.69	0.70
Sarcasm	0.62	0.66	0.64
Vulgarity	**0.75**	0.74	**0.75**
Abuse	0.77	0.78	0.76

6. Discussion

Key findings:

- Preprocessing improved sentiment and abuse detection by handling noisy OCR.
- CNNs captured lexical sarcasm cues better than linear models.
- Multimodal fusion was critical for vulgarity detection, where offensiveness was primarily visual.
- Lexicon-informed features improved recall for rare abusive expressions. The offensive term replacement was implemented using a simple dictionary-based lookup, where any exact match of a term in the `hindi-offensive-words-original.json` was replaced with a neutral placeholder. This ensures that the model learns the context of the meme without explicitly learning to classify based on individual vulgar terms.

7. Conclusion and Future Work

We presented **Team FAST's** robust multimodal system for HASOC 2025 Hindi memes. By combining tailored preprocessing, classical machine learning, and neural architectures, we achieved competitive performance across all subtasks, with the best Macro-F1 score of **0.75** for vulgarity detection. Future directions could include:

- End-to-end multimodal transformers (mBERT, CLIP).
- Larger Hindi-Hinglish pretrained embeddings.
- Fine-grained abuse target classification.

Code and Resources

All code, stopword lists, offensive word dictionaries, and models are available at: https://github.com/fatimahansari/hindi-HASOC-2025

Declaration on Generative AI

During the preparation of this work, the author(s) used Grammarly in order to: Grammar and spelling check. After using these tool(s)/service(s), the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content

References

[1] Koyel Ghosh and Mithun Das and Mwnthai Narzary and Saptarshi Saha and Shubhankar Barman and Animesh Mukherjee and Sandip Modha and Debasis Ganguly and Utpal Garain and Sylvia Jaki and Thomas Mandl, Overview of the HASOC Track at FIRE 2025: Abusive Meme Identification — Shadows Behind the Laughter, in: K. Ghosh, T. Mandl, S. Pal (Eds.), Forum for Information Retrieval Evaluation (Working Notes) (FIRE 2025) December 17-20, Varanasi , India, CEUR-WS.org, 2025.

[2] T. Ranasinghe, K. Ghosh, A. S. Pal, A. Senapati, A. E. Dmonte, M. Zampieri, S. Modha, S. Satapara, Overview of the hasoc subtracks at fire 2023: Hate speech and offensive content identification in assamese, bengali, bodo, gujarati and sinhala, in: Proceedings of the 15th annual meeting of the forum for information retrieval evaluation, 2023, pp. 13–15.

[3] K. Ghosh, D. A. Senapati, Hate speech detection: a comparison of mono and multilingual transformer model with cross-language evaluation, in: S. Dita, A. Trillanes, R. I. Lucas (Eds.), Proceedings of the 36th Pacific Asia Conference on Language, Information and Computation, Association for Computational Linguistics, Manila, Philippines, 2022, pp. 853–865. URL: https://aclanthology.org/2022.paclic-1.94/.

[4] M. Das, S. Banerjee, A. Mukherjee, Data bootstrapping approaches to improve low resource abusive language detection for indic languages, in: Proceedings of the 33rd ACM conference on hypertext and social media, 2022, pp. 32–42.